Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies

Authors:
Francesco Cicconardi[a,1], Edoardo Milanetti[b,c], Érika C. Pinheiro de Castro[d], Anyi Mazo-Vargas[e], Steven M. Van Belleghem[f,g], Angelo Alberto Ruggieri[f], Pasi Rastas[h], Joseph Hanly[i,j], Elizabeth Evans[f], Chris D Jiggins[d], W Owen McMillan[j], Riccardo Papa[f,k], Daniele di Marino[l,m], Arnaud Martin[i], Stephen H Montgomery[a,j,1]

Affiliations:
[a] School of Biological Sciences, Bristol University, UK.
[b] Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00185, Rome, Italy
[c] Center for Life Nano- & Neuro-Science, Italian Institute of Technology, Viale Regina Elena 291, 00161 Rome, Italy
[d] Department of Zoology, University of Cambridge, Cambridge, United Kingdom.
[e] Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853
[f] Department of Biology, University of Puerto Rico, Rio Piedras, Puerto Rico.
[g] Ecology, Evolution and Conservation Biology, Biology Department, KU Leuven, Leuven, Belgium.
[h] Institute of Biotechnology, University of Helsinki, Helsinki, Finland.
[i] Department of Biological Sciences, The George Washington University, Washington, DC 20052, United States
[j] Smithsonian Tropical Research Institute, Panama.
[k] Molecular Sciences and Research Center, University of Puerto Rico, San Juan, PR.
[l] Department of Life and Environmental Science, New York-Marche Structural Biology Center (NY-MaSBiC), Polytechnic University of Marche, Ancona, Italy.

[1] Corresponding author: francicco@gmail.com; s.montgomery@bristol.ac.uk

Short title: *Tribe-wide Heliconiini genomics*
Teaser: *Dense sampling reveals the genomic basis of key innovations in an enigmatic tribe of butterflies.*

Abstract

Heliconius butterflies, a speciose genus of Müllerian mimics, represent a classic example of an adaptive radiation involving a range of derived dietary, life history, physiological and neural traits. However, key lineages within the genus, and across the broader Heliconiini tribe, lack genomic resources, contrasting our understanding of how adaptive and neutral processes shaped genome evolution across their radiation. Here, we build new, highly-contiguous genome assemblies for nine new Heliconiini, reference-assembled genomes for 29 species, and improve 10 existing assemblies, to provide a major new dataset of annotated genomes for 63 species, including 58 species within the Heliconiini tribe. We provide a robust, dated heliconiine phylogeny, identify major patterns of introgression, explore the evolution of genome size, content, and the genomic basis of key innovations in this enigmatic group for the first time. We illustrate how dense genomic sampling improves our resolution of gene-phenotype links, and our understanding of how genomes evolve.

A central goal of evolutionary biology is to understand how biodiversity is generated, maintained, and how organisms' interactions drive the diversity of natural communities. Periods of rapid diversification are often associated with the colonisation of new evolutionary niches or the exploitation of new resources (*1*). The evolution of key innovations, such as physiological adaptation to food resources, or new morphological traits, can enable these ecological shifts, and play critical roles in adaptive radiations (*2*). Identifying and understanding the molecular and genetic basis of such key innovations is now a realistic goal (*3*, *4*), and can provide explicit links between genetic changes, natural selection and speciation, in the context of wider patterns of genomic divergence. Heliconiini, a Neotropical tribe of Nymphalid butterflies, comprised of ~80 species and ~400 subspecies, have become a key system with which to explore the biology of speciation (*5*–*8*). In particular, the rapid radiation of the genus *Heliconius*, and their diversity of colour patterns, have become a case study in how genomic approaches can improve our understanding of the genetic architecture of adaptive traits, and the accumulation of reproductive isolation with ongoing gene flow (*7*). Heliconiini also exhibit key innovations including the tribe-wide restricted use of Passifloraceae as larval hostplants – their antagonist coevolutionary partners that can provide them with cyanogenic glucosides for chemical protection (*9*). Within them, species of the genus *Heliconius* are also the only lepidoptera to actively collect and digest pollen as adults, which is associated with major shifts in reproductive lifespan (*10*), and behavioural and neural elaboration (*11*). As such, the availability of tribe-wide genomic resources would represent a major resource to explore the biology of an enigmatic case study in adaptive diversification.

Here, we provide such a resource by sequencing, assembling and using a combinatorial approach to maximise methodological approaches, with a unified cross-species annotation to remove possible species-biases previously unrepresented in available data. Combined with improved, both in terms of assembly contiguity and gene annotation, of already available resources, we generate a new and comprehensive dated phylogeny for Heliconiini, and explore structural and adaptive aspects of genome evolution of ecological innovation across the radiation and their key transitions. Our genomic dataset comprises ~75% of all the species in the Heliconiini tribe, the most comprehensive effort to densely sample genomic data in any insect tribe.

Improved Resolution of Phylogenetic Relationships and Signatures of Introgression

A species-level phylogeny was determined with a maximum-likelihood (ML) analysis, and used to estimate divergence dates (Fig. 1C, fig. S16, and table S3). Although the topology is widely consistent with previously inferred phylogenetic relationships (*8*), we identify differences within some

Heliconius clades, where the Silvaniform and Melpomene clades are now paraphyletic, and among other genera of Heliconiini, with *Podothricha telesiphe* and *Dryas. iulia* now sister lineages, outgrouped by *Dryadula phaetusa* and *Philaethria. dido*. The estimated divergence times show that the subfamily Heliconiinae originated ~45.3 million years (Mya) (95% CI: 35.9-55.5), with the last common ancestors of *Eueides* and *Heliconius* dating to ~11.1 Mya (95% CI: 7.3-12.1) and 9.6 Mya (95% CI: 8.8-13.8), respectively. Interestingly, deeper branches of the phylogeny are characterized by high molecular substitution rates (Fig. 1C and table S3), indicating a series of bursts in evolutionary rate at

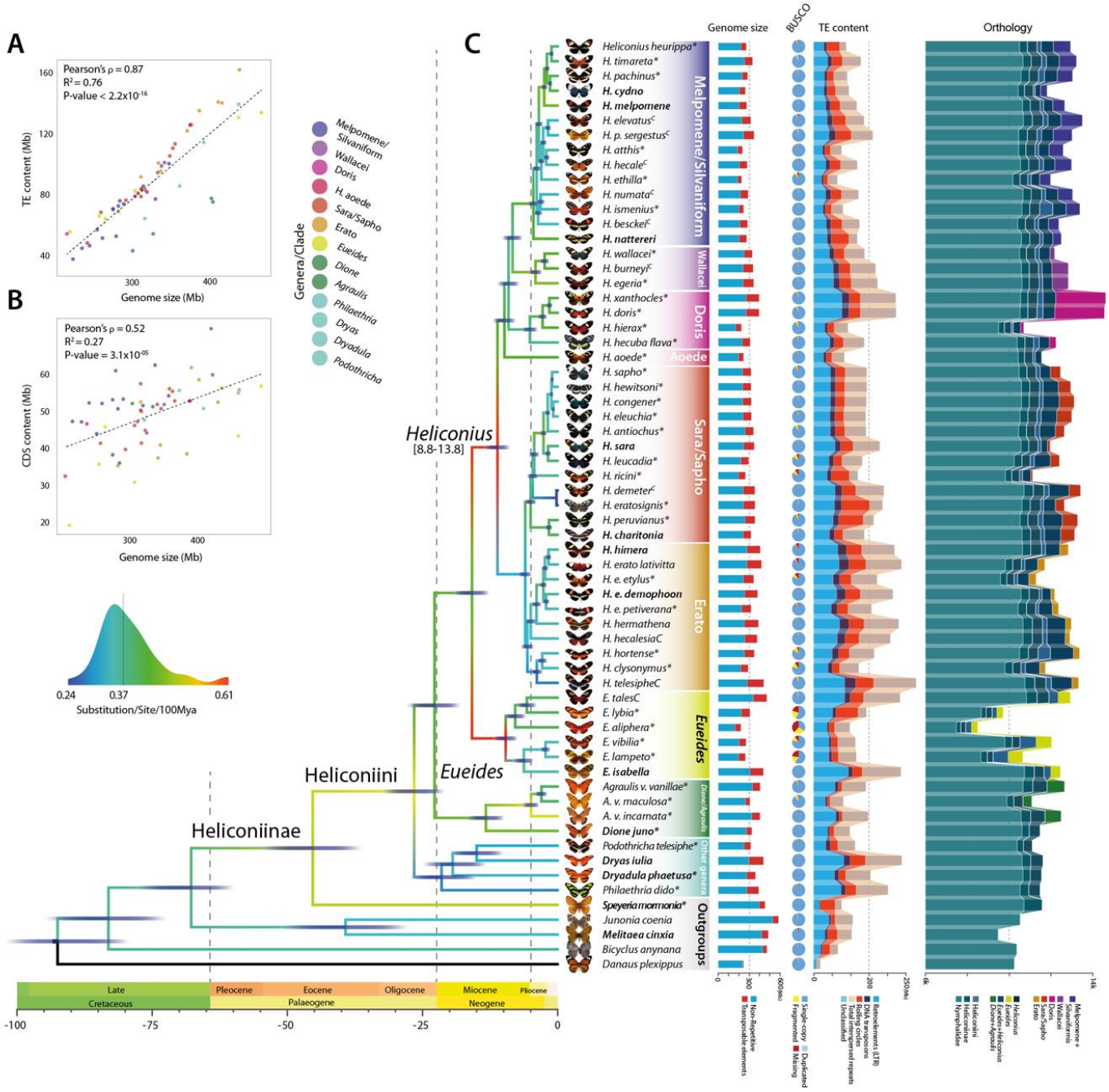

Fig.1 | Phylogenetic, genomic, and proteomic comparisons among 63 Nymphalid butterfly species. (**A**, **B**) The contribution of TEs and CDS to genome size variation across Heliconiinae, respectively. (**C**) From left to right: *i*) the dated species phylogeny built from the

concatenated single-copy orthologous groups (scOGs) from all sequenced Heliconiinae and outgroups, using a combination of Maximum Likelihood and Bayesian Inference. The branch colour represents the number of substitutions per site per 100 Mya of that specific branch. Species names in bold indicate the species with chromosome- or sub-chromosome-level assemblies, asterisks indicate genomes assembled in this study, [C] curated assemblies; *ii*) genome assembly size, in red the TE fractions; *iii*) BUSCO profiles for each species. Blue indicates the fraction of complete single-copy genes; *iv*) bar plots show total gene counts partitioned according to their orthology profiles, from Nymphalids to lineage-restricted and clade-specific genes.

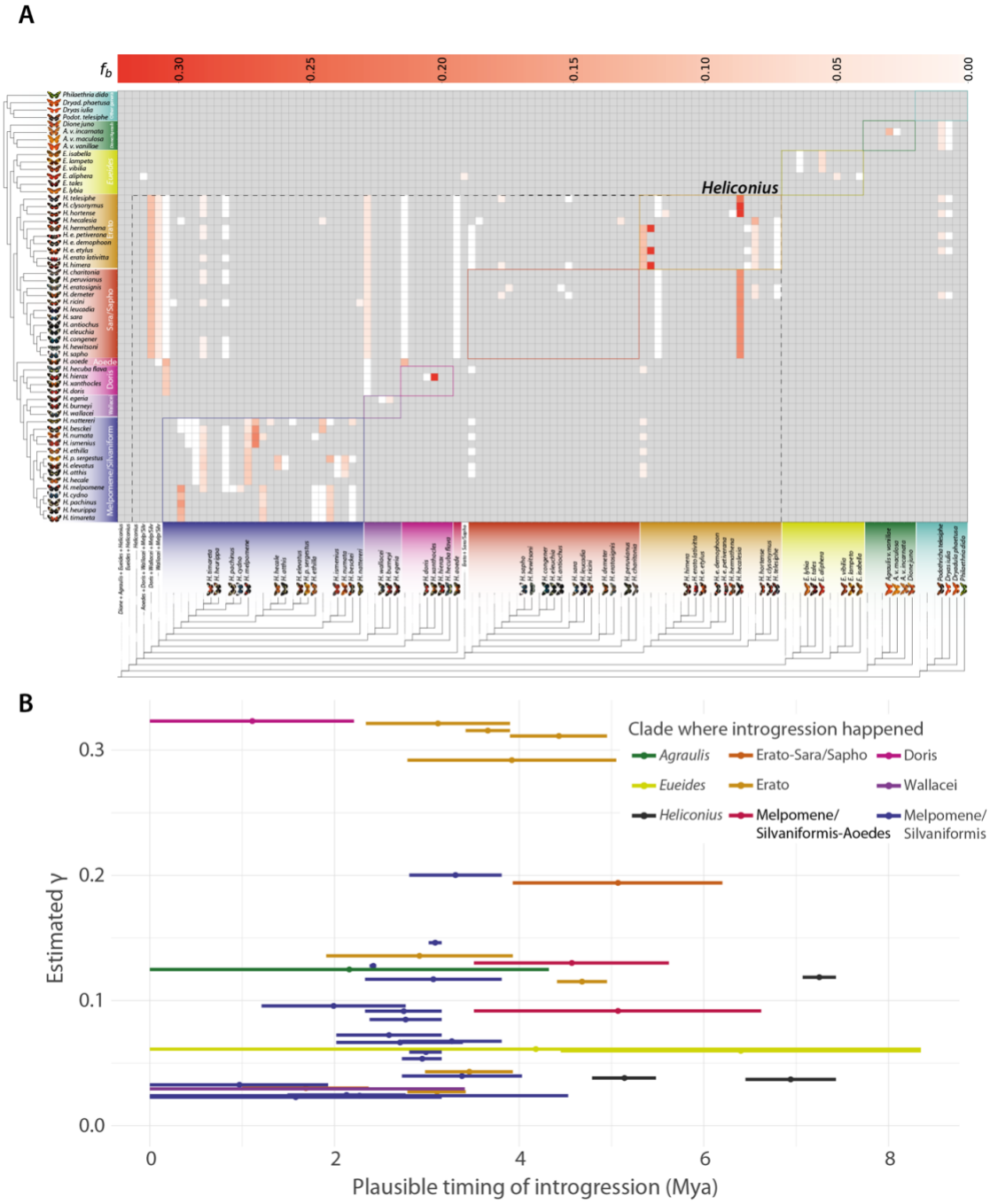

Fig. 2 | Patterns of introgression inferred for the Heliconiinae clade. (**A**) The matrix shows inferred introgression proportions as estimated from scOG gene trees in the introgressed species pairs, and then mapped to internal branches using the f-branch method. The

expanded tree at the bottom of the matrix shows both the terminal and ancestral branches. (**B**) Each segment indicates the confidence interval of the estimated introgression event (triplet support ratio > 10%). The circles indicate the average date.

the base of the radiation. Of critical importance for understanding the adaptive evolution of *Heliconius*, and the set of traits linked to pollen feeding, is the position of the Aoede clade, which do not pollen feed but are grouped within *Heliconius* in our topology (Fig. 1C, and fig. S16). This position is strongly supported by bootstrap and concordance values (Fig. 1C and fig. S16), with the majority of quartets supporting the focal branch (fig. S18 and fig S19; QC = 0.9).

Heliconius have also become a key lineage for exploring of the impact of gene flow and hybridization on adaptive divergence (*5*, *7*), we, therefore, revisited this topic with our improved taxonomical range, adopting two new methodological approaches: the discordant-count test (DCT) and the branch-length test (BLT). Both tests reveal a lack of gene flow between basal *Heliconius* nodes, including the *H. aoede* split, but do identify a number of introgression events within major clades of Heliconiini (Fig. 2, fig S20 and table S4). The greatest number of introgression events were detected within *Heliconius*, specifically between the most recent common ancestors (MRCAs) of Erato + Sara/Sapho clade, the Doris + Wallacei + Silvaniform + Melpomene clade, and within the Erato clade. Interestingly, the Sara/Sapho clade shows very low rates of introgression, potentially reflecting a greater tendency for monandrous females (*12*) in this clade where males often mate with females as they eclose from the pupae (*13*). The estimated fraction of introgressed genome mostly varies between 0.02 γ to 0.15 γ, with a peak around 0.30 within the Erato clade (range of average γ estimates = 0.023–0.323). Most introgression events also occurred in a restricted time frame within the last 5 Mya (Fig. 2B), and no significant relationship was found between the midpoint estimate of the timing of introgression and the estimated γ (Fig. 2B), indicating that the fraction of a genome that is introgressed within *Heliconius* does not depend on the timing of those introgression events.

Evolution of Genome Size and Content

Variation in genome size can be formalised in an "accordion" model (*14*) where genomes gain, lose, or maintain its size in equilibrium in each species, due to a balance between expansions in transposable elements (TEs) and large segmental deletions. By reconstructing ancestral genome sizes at each node in the Heliconiini phylogeny, we found that the MRCA of Heliconiinae experienced a constant reduction for all major internal branches (*e.g.*, from ~367Mb for Heliconiini branch to ~308Mb for Wallacei + Silvaniform + Melpomene branch). At the same time, other branches leading to *Philaethria*, *Dryadula*, *Dryas* and *Podothricha*, and the Erato, Doris and Wallacei clades within the genus *Heliconius,* had independent expansions. Strikingly, *H. aoede* shows a loss of about 68 Mb, a fifth of its genome size (~22%) from its ancestral node (Fig. 3).

For the two most speciose clades, *Heliconius* and *Eueides*, we tested whether different genomic compartments (CDS, introns, 5′-UTR, and 3′-UTR) contribute to the change in genome size equally by calculating CONACC scores and assessing departures from neutrality (Fig. 3B). Between the two genera, we identified an enrichment for higher CONACC scores in *Heliconius* for CDS and introns, compared to the same compartments in *Eueides*, a trend that is inverted for the two UTR regions (Wilcoxon rank-sum test 'two-sides' *P* value < 2.2×10^{-16}). This suggests an increased tendency for clade-specific selection, also confirmed by the adopted fast-unconstrained Bayesian approximation method (FUBAR), which showed more sites under purifying selection than under positive selection, by several orders of magnitude (Fig. S21). *Heliconius* also has 2.5x more sites under purifying selection per codon than *Eueides*, suggesting that the higher CONACC scores in *Heliconius* are likely due to higher degrees of purifying selection.

We next explored the relationship between transposable elements (TEs) and genome size, and their effect on gene architecture. We found that larger genomes

tend to have a distribution of intron length skewed towards longer introns (fig. S16a), with a positive correlation between median intron length and total TE content (fig. S22; Pearson's ρ=0.72; R^2=0.51). Long introns also accumulate significantly more TEs then expected by their size (fig. S23; Wilcoxon rank-sum test P value = 2.13×10^{-13}), with the effect of changing gene structure more than gene density (fig. S22 and fig. S24). This suggests that selection may have favoured a reduction in TEs in intergenic regions, perhaps to avoid the disruption of regulatory elements (*15*), consistent with TEs largely accumulating in the tails of chromosomes (*16*). Although intron size varies significantly, the rate of gain/loss of introns, and the intron retention from the MRCA of Nymphalids, shows a relatively stable dynamic over the last 50 Mya in Heliconiinae, with no significant shift among species, and ~7% of ancestral intron sites retained across species (Fig. 4C). A similar pattern was reported in *Bombus* (*17*), but our results differ from drosophilids and anophelines, which show significantly higher intron turnover rates (*18*).

Expansion and Contraction of Gene Content

Ortholog group (OG) size evolution was modelled with CAFE for 10,361 OGs using the 52 most complete genomes (BUSCO score ≥ 90%). The analysis identified 656 OGs that vary significantly across the phylogeny. The estimated gene turnover (λ) was of 0.006/gene gain-loss/Mya. This is relatively high compared with rates for *Bombus* (λ = 0.004) and anopheline species (λ = 0.003) (*17*, *18*), but similar to drosophilids (λ = 0.006) (*19*). The base of

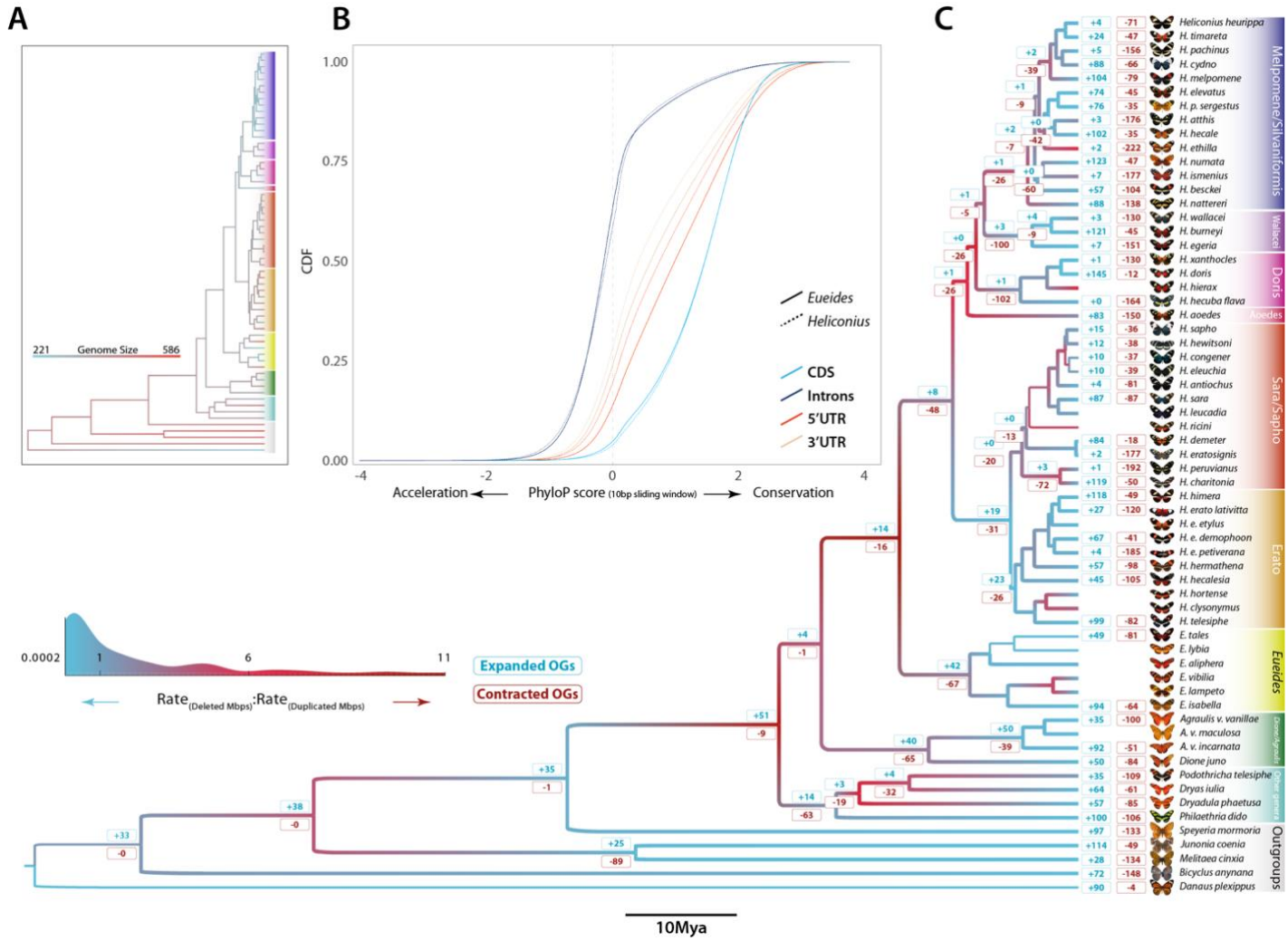

Fig. 3 | Genomic dynamics, acceleration/conservation rates and ortholog copy number evolution. (**A**) Ancestral genome size reconstruction of Nymphalids inferred by ML approach. (**B**) Cumulative distribution functions (CDFs) of scores for selected annotation classes (CDS, introns, 5' and 3' UTRs) as computed by the subtree scores for *Eueides* and *Heliconius* clades. PhyloP scores at sites of different annotation classes, based on the LRT method and multiple whole 63-species genome alignment. Positive scores indicate conservation, and negative scores indicate acceleration (CONACC mode) in a 10-bp sliding window. (**C**) Branch colours indicate the ratio between the rate of duplications (duplicated Mb/Mya per branch) and deletions (deleted Mb/Mya per branch) across the Nymphalid phylogeny. Numbers at nodes correspond to the amount of expanded (blue) and contracted (red) ortholog groups (values are shown for main branches and most complete genomes, see Methods).

the phylogeny showed relatively strong OG expansions, with few contractions, followed by stasis. While *Dione* + *Agraulis* and *Eueides* stems have similar proportions of expanded/contracted OGs, *Heliconius* shows 48 contracted OGs but only eight expanded OGs (Fig. 3C, table S5) suggesting the phenotypic innovations that occurred in this branch were not due to widespread gene duplication.

Several OGs were identified that expanded multiple times across the phylogeny. Cytochrome P450 genes expanded in the common ancestor of the subfamily Heliconiinae, the tribe Heliconiini, the *Dione* + *Agraulis* stem, and within the genus *Heliconius* in the Erato group and Silvaniform/Melpomene stems. Insect Cytochrome P450s have roles in the detoxification of specialized metabolites, hormone biosynthesis/signalling, and biosynthesis of cyanogenic glucosides in heliconiine butterflies, which form the basis of their chemical defence (*20*). Notably, a range of diet related OGs are also highlighted: Glucose transporters and Trypsins expanded several times in

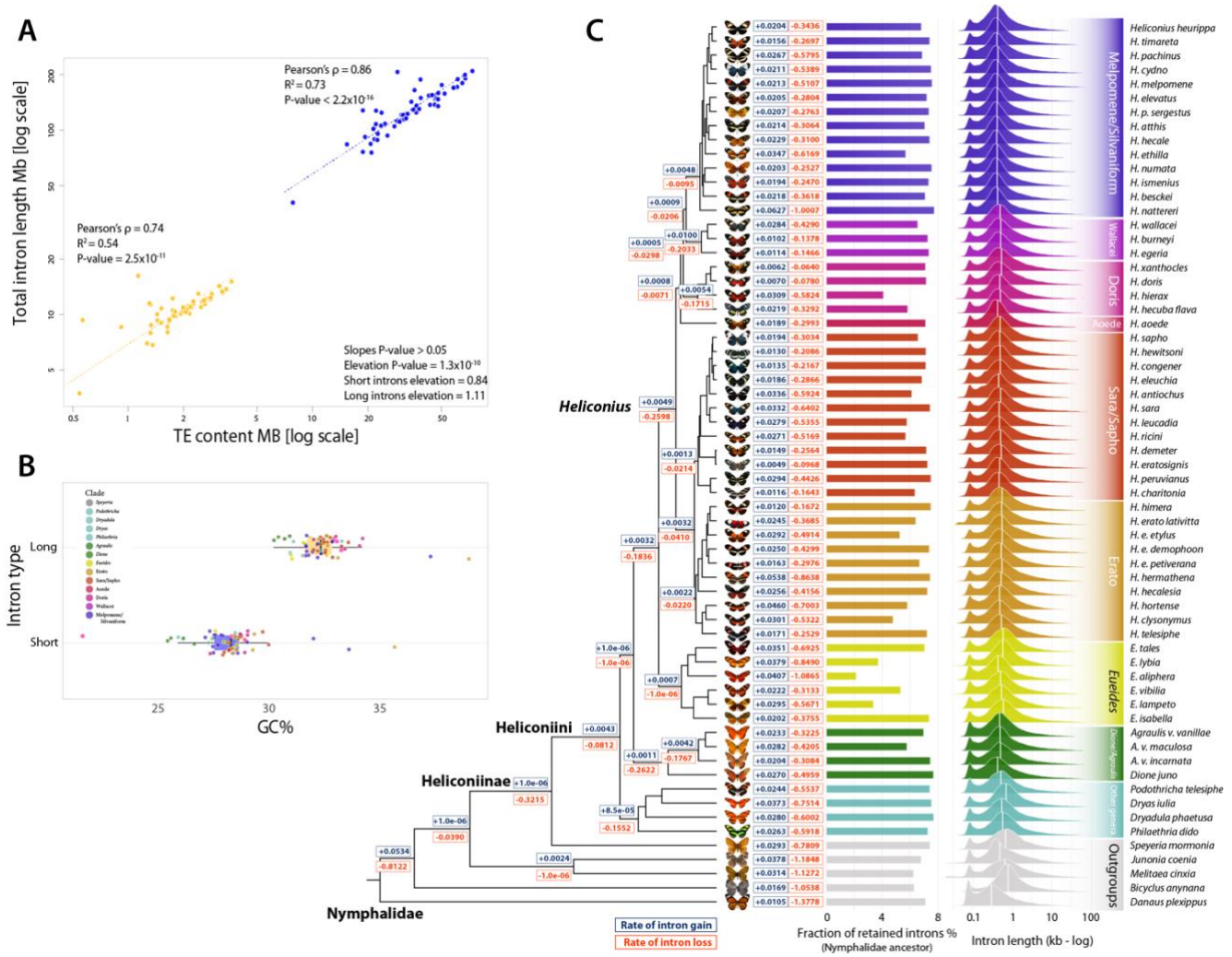

Fig. 4 | Composition and intronic evolution. (**A**) Log-log scatter plot showing high correlation between intron length and TE abundance. The significantly higher elevation of longer introns compared with short introns, indicates that large introns are more affected by transposable elements (TEs); short intron elevation = 0.84; long intron elevation = 1.11; *P* value = 1.3×10^{-10}. (**B**) Box plot of GC content and intron size (Wilcoxon rank-sum test *P* value < 2.2×10^{-16}). (**C**) Rates of intron gain (blue) and loss (red) across Nymphalid phylogeny, and fraction of retained introns from the Nymphalid ancestor. On the far left, the log scale distributions of intron lengths, white vertical bars indicate the median values.

Heliconiinae, Heliconiini, *Eueides*, and the Silvaniform/Melpomene stem. Although glucose transporters play an important role in energetic metabolism in all animals, in phytophagous insects they are also hypothesised to be involved in the sequestration and detoxification of specialized metabolites from plants (*21*). Lipases, expanded in the Sara/Sapho + Erato stem and Silvaniform/Melpomene clades, are key enzymes in energetic metabolism and in pheromone biosynthesis (*22*). These results therefore suggest potential links to the chemical ecology of the Heliconiini. In addition, the *Heliconius* stem had an expansion of *methuselah*-like, a G-protein coupled receptor, involved in oxidative stress response, metabolic regulation, and lifespan (*23*). *Esterase P*, and a *juvenile hormone acid methyltransferase* (*jhamt*) also expanded three

times in *Heliconius* stem, providing candidates for derived life history traits.

We analysed 57 gene families (GF) in more detail, including GFs with a range of biological functions (table S6). We used measures of "phylogenetic instability" and the gene turnover rate (λ, CAFE), to explore their dynamics, and found that the average instability score was 37.45 while the average λ is 0.005, with the number of OGs per family positively correlated with λ (Pearson's ρ = 0.42). Sodium/calcium exchanger proteins, and Hemocyanin superfamily show the highest instability and turnover rates (fig. S33). This analysis also identifies GFs which expanded in key periods of heliconiine diversification, including Hemocyanins, Lipases, Trypsins and Sugar transporters and the Major Facilitator Superfamily (Fig. 5A). The most interesting are CYP303A1-like gene (fig. S34), a highly conserved protein in insects which has a pivotal role in embryonic development and adult moulting (*24*), and two Hemocyanins: a hexamerin storage homolog expanded in the *Dione* + *Agraulis* + *Eueides* + *Heliconius* clade, and the arylphorin homolog, in *Eueides* and *Heliconius*. These hexamerins function as storage proteins, providing amino acids and energy for non-feeding periods, such as molting and pupation, and may also transport hormones (*25*). Hemocyanins are also among several GFs with evidence of divergent selection regimes (*ω*) between *Eueides* and *Heliconius,* alongside Trypsins, Protein kinases, P450s, Sugar transporters, Ion and ABC transporters (Fig. 5B). Curiously, a contraction in the Hemocyanin superfamily was only observed in *H. aoede*, the only *Heliconius* species not to feed on pollen in our data, marking hexamerins as a potential mechanistic link to the divergent strategies for nitrogen storage in pollen-feeding *Heliconius* (Fig. 5A).

Selection Across Heliconiini Radiation

Selection regimes shaping Heliconiinae coding genes were further investigated using the adaptive branch-site random effects likelihood (aBSREL) method. When classified according to their phylogenetic attribution (*i.e.,* where they appeared throughout the phylogeny), single-copy orthologous groups (scOGs) show a trend towards an increase of purifying selection from young to older genes (Fig. 5C), suggesting that genes become more stable with time, probably reflecting increased functional importance. The signature of diversifying positive selection was assessed on the five main branches of the Heliconiinae radiation. From the base of Heliconiini to *Dione* + *Agraulis* + *Eueides* + *Heliconius*, *Eueides* + *Heliconius*, *Eueides*, to the *Heliconius* stems. Overall, the Heliconiini branch evolved under the strongest selection, followed by the *Eueides* and *Heliconius* branches, and finally by the *Eueides* + *Heliconius* branch (Fig. 5D). The number of genes with a signal of diversifying positive selection varies between branches, with the *Dione* + *Agraulis* + *Eueides* + *Heliconius* and *Heliconius* stems having the highest number of enriched biological processes (BPs), followed by Heliconiini and *Eueides* stem, and *Eueides* + *Heliconius*. A notably high proportion of branches are enriched for BPs relating to neuronal development and cellular functions, including the regulation of hippo signalling, stem cell differentiation and cell-cell adhesion, and genes associated with asymmetric division. Using a network-based approach, which integrates both primary and predicted interactions to predict gene function we examined connections between selected genes. Although the amount of network interactions shows a significant degree of connectivity (absolute number of interactions) in the branches leading to *Dione* + *Agraulis* + *Eueides* + *Heliconius* (834 interactions), *Eueides* + *Heliconius* (627), *Eueides* (531), Heliconiini (410), and *Heliconius* (320), the network density shows a different picture, with *Heliconius* having a markedly higher density (the portion of the potential connections in a network that are actual connections) with ~0.3 versus ~0.2 for the other networks, in the case of BP networks (fig. S35, fig. S36, and table S9). The enriched molecular functions (MFs) in this densely connected *Heliconius* network are characterised by BPs related to response to DNA damage/repair, neuroblast division, and neural precursor cell proliferation, glial cell development, cell-cell junction assembly, asymmetric stem cell division. This concentration of neurogenesis-related functions differs from enrichment in other networks, which appear more variable. Finally, we note multiple genes that show a signature of diversifying positive selection on more than one branch. One of them is the *Notch* homolog, an essential signalling protein with major roles in developmental processes of the central and peripheral

nervous system (*26*). Notch regulates neuroblast self-renewal, identity and proliferation in larval brains, and it is involved in the maintenance of type II neuroblast self-renewal and identity (*27*).

***Candidate Genes for the Derived Traits of* Heliconius.**

Within the Heliconiinae, *Heliconius* displays a number of divergent traits and innovations (*10*). Here, we exemplify how our results reveal new biological insights into these traits, focusing on mushroom body expansion and the enzymatic processes associated with breaking down pollen walls to aid their digestion. Mushroom bodies are paired organs within the central brain that receive visual and/or olfactory information, and play a pivotal role in learning and memory (*28*). These structures show huge variation across Heliconiini with major expansions in the *Heliconius* stem (*11*, *29*), but the molecular mechanisms underpinning these events are unknown. Our selection analyses highlight pathways that could regulate neural proliferation. These include the Hippo signalling pathway, which regulates cell growth and proliferation of neural stem cells and neuroblast quiescence (*30*). Multiple and repeated signs of diversifying selection are identified on genes related to the Hippo signalling pathway, including *Focal adhesion kinase* (*Fak*)*, lethal (2) giant larvae* (*lgl*)*, Sarcolemma associated protein* (*Slmap*)*,*

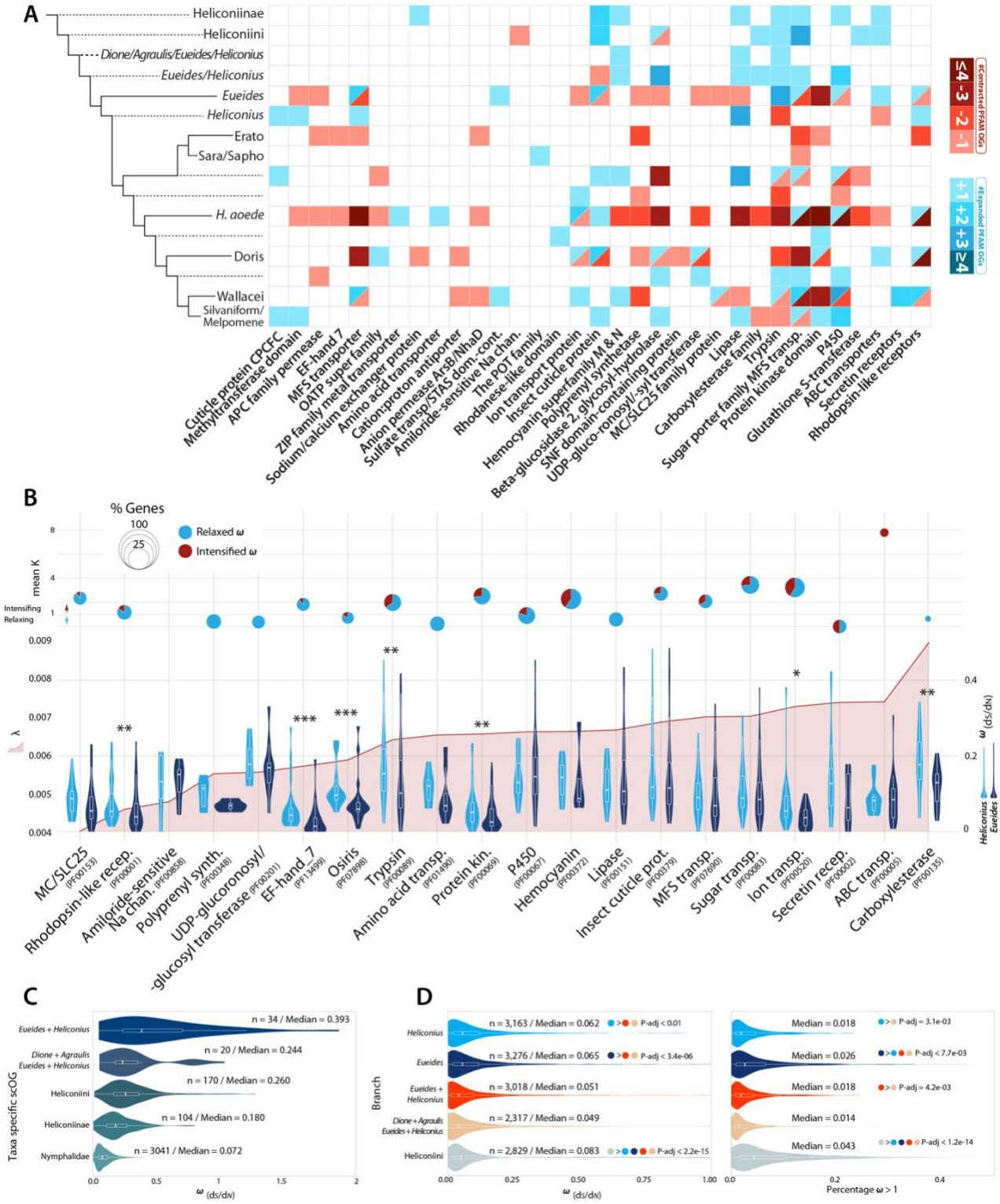

Fig. 5 | Differential evolutionary rates gene families and scOGs across Heliconiini butterflies. (**A**) Heatmap showing the different expansions and contractions in multiple gene families. Several gene families have been contracted in *H. aoede*. (**B**) Plots showing different evolutionary features of some of the analysed gene families (minimum of 3 genes in *Eueides spp*. and *Heliconius spp*.). At the top section, dynamic pie charts showing mean *K* value (selection intensifier parameter). Values below one indicates a relaxation, while above one

indicates intensification towards diversifying positive selection. The size of pie charts indicates the fraction of genes under intensification (red) and relaxation (blue), and it is scaled according to the proportion of genes for which K was significantly different from H_0 (No difference) (see Methods). For different gene families the panel below shows the gene turnover rate (λ) (left y-axis); right y-axis shows the distributions of mean ω for near-scOGs (see Methods) in *Eueides* and *Heliconius* (right y-axis). Asterisks indicate significant shifts between *Eueides* and *Heliconius* (Wilcoxon rank-sum tests). (**C**) Violin plot showing the distributions of mean ω rates (d_N/d_S) in scOGs according to their lineage-specificity. (**D**) Distribution of mean ω rates (left) for scOGs on six branches of Heliconiini, and the proportion of genes for which ω is higher than one (right).

and *Akt kinase* (*Akt*), on the *Dione* + *Agraulis* + *Eueides* + *Heliconius* stem, which regulate cell polarity, asymmetric division and cell proliferation (*31*), and two other genes, *Moesin* (*Moe*) and *F-box and leucine-rich repeat protein 7* (*Fbxl7*), in the *Eueides* + *Heliconius* stem. Moe drives cortical remodelling of dividing neuroblasts (*32*), while Fbxl7 affects Hippo signalling pathway activity (*33*). Finally, Ctr9, dachsous (ds), *falafel (flfl)*, and locomotion defects (loco) were identified at the *Heliconius* stem (table S8). Ctr9 is involved in the proliferation and differentiation of the central nervous system (*34*), Flfl is required for asymmetric division of neuroblasts, cell polarity and neurogenesis in mushroom bodies (*35*), Ds is a cadherin that interacts with the Hippo signalling pathway (*36*), and loco is an activator of glial cell fate, essential cells in efficiently operating nervous systems (*37*).

Despite being a keystone innovation in *Heliconius*, little is known about the mechanism underpinning pollen feeding. Saliva probably has an important role in the external, enzymatic digestion of the pollen wall (*10*). The leading candidates for these enzymes are serine proteases, homologs of the silkworm cocoonase (*38*, *39*), which digests the cocoon during eclosion (*39*). Because butterflies do not produce a cocoon, it has been proposed that the duplications of cocoonase orthologs may have been co-opted to digest pollen (*38*). Given our order of magnitude larger sample, we re-evaluated the evolution of these genes by reassessing the evolutionary history of this gene family, and evidence of gain-of-function. We identified 233 cocoonase loci (table S10) across all Heliconiinae, and found that the duplications not only predate the split between *Heliconius* and *Eueides,* but affect the whole Heliconiini tribe and its outgroup *S. mormonia* (Fig. 6A). All species have at least four copies, located at the minus strand of chromosome 15 with remarkable synteny (Fig. 6B and table S11). Substrate, cleavage, and active sites of the functional domain shows very high conservation throughout the dataset. Three independent tandem duplications from the same original copy are very likely responsible of the emergence of *Coc1A, Coc1B, Coc2* and *Coc3* (Fig. 6A). High level of purifying selection is detected across the four OGs (Fig. 6A). A scan of all internal branches for signs of diversifying positive selection shows that the branches of *D. juno Coc2* in-paralogs; the stems of all *Heliconius Coc1A* and *Coc1B*, and the two branches of the Silvaniform/Melpomene *Coc2* out-paralog, show signs of positive selection. Two of these events involve the loci of the non-pollen feeding *H. aoede*. We therefore tested if these loci show signs of relaxation in this species. Surprisingly, while no significant differences were detected for *Coc1B*, an intensification of selection was detected for *Coc1A* (K = 1.41; P-value = 0.001). To gain more insight into a gain-of-function hypothesis, we modelled the 3D structure of the full-length protein sequences, a trypsin-like serine protease composed of two folded beta barrels connected by a long loop positioned at the back of the active cleft (*38*), and, by adopting a new graph-based theory approach, we inferred the key residues driving the structural differences among loci. Notably, the methodology clustered all the structures into four groups, consistent with the phylogenetic analysis, plus a fifth group for the Melpomene/Silvaniform *Coc2* sub-clade, which evolved under diversifying positive selection. We found that seven residues drive the overall clustering, and these lie in three regions of the 3D structures, corresponding to three loops in regions highly exposed to the solvent (Fig 6D). The structural alignment of the predicted cocoonases with the x-ray

structures of several homologous human serine proteases (fig. S40), shows that the two largest loops (pos: 217-217 and pos: 119-122) corresponds to highly flexible regions in the experimental structures (*i.e.*, B-factor), in contrast with the shorter third loop (pos: 68-71), which is in turn analogous to a region with higher stability. These analyses suggest that the duplicated genes might have gained the capacity to bind and process different substrates by changing their flexibility throughout the radiation of Heliconiini. This is consistent with the hypothesis that, in order to obtain a gain-of-function and to give rise to new interactions, a protein needs to change few sites in intrinsically disordered regions (*40*). Our combined results present a more complex story than previously described, and both the high copy number variation and patterns of selection within Heliconiinae appear inconsistent with these genes playing a critical role in the evolution of pollen feeding.

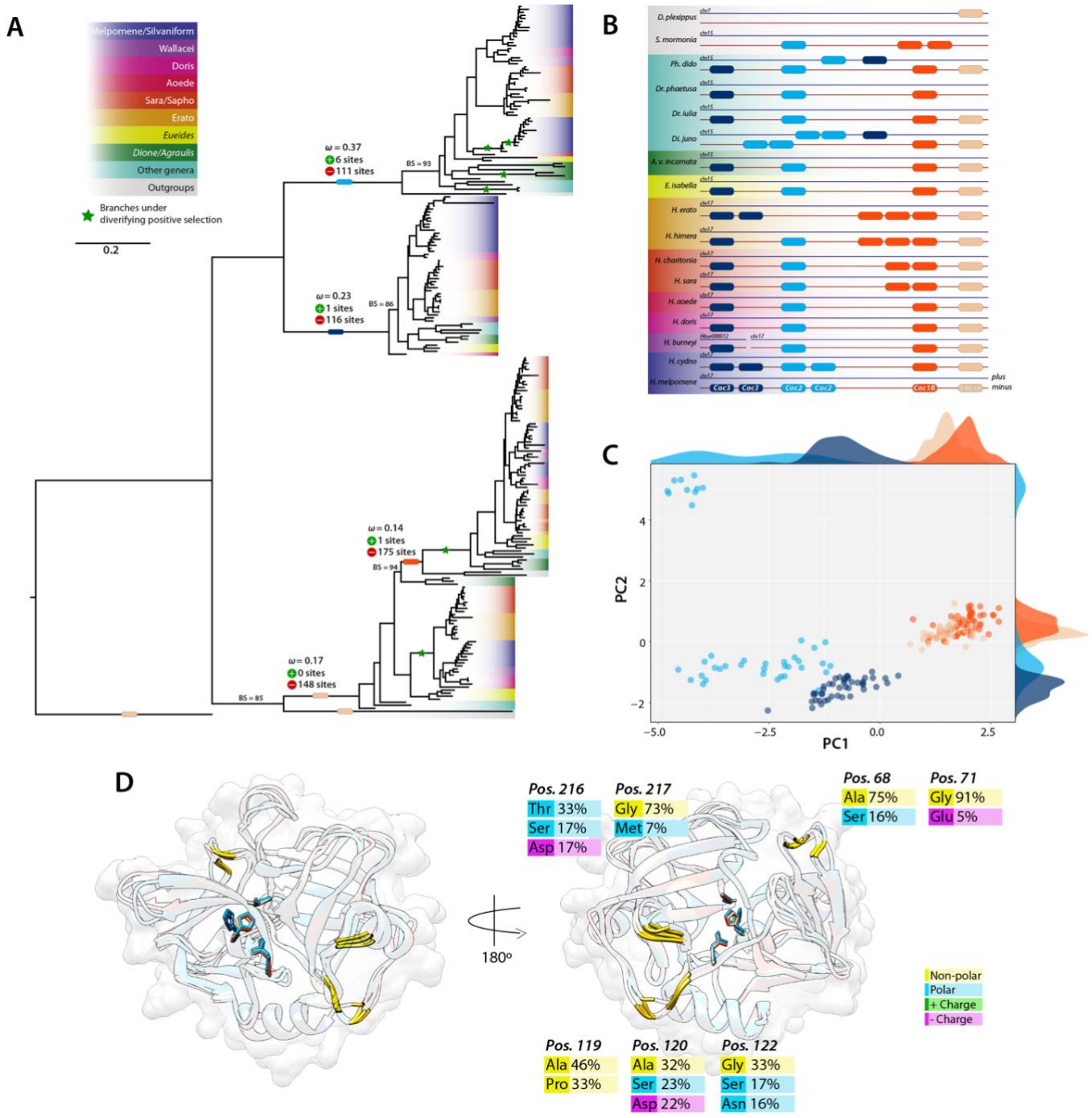

Fig. 6 | Cocoonase evolution and structural divergence across Heliconiini. (**A**) Maximum likelihood phylogeny of nucleotide sequences of the four cocoonase loci across Heliconiinae. Coloured blocks on branches show the stem for each locus. The green star indicates branch under diversifying positive selection. Bootstrap (BS) values for main branches are listed for those with values below 95 (**B**) Synteny map of the different loci for the genomes with highest contiguity. (**C**) PCA of the network-based analysis of 181 predicted protein models. (**D**) Structural alignment of the closest sequence to the centroid of each of the five clusters of the PCA (coloured ribbons). For each of structure the three active sites of the active cleft are depicted as sticks, while in gold the seven identified positions the best

explains the clustering of the PCA. On the left the same structure rotated 180º. For each of the positions the most frequent amino acids are shown with their respective frequency in the alignment.

Conclusions

We have curated available genomic data and new reference genomes to build a tribe-wide dataset for Heliconiini butterflies. The density and breadth of this dataset allowed us to resolve ambiguities in the dated phylogeny, identify major introgression events, and explore the evolution of genome size and content. The characterization of all these genomic features, ultimately allowed us to study the molecular architecture of key innovations in this enigmatic group of butterflies. In doing so, we provide genome-wide insights into potential molecular mechanisms underpinning major innovations in Heliconiini, including expansions in gene families and OGs linked to biochemical processes relevant to cyanogenic defences, dietary shifts, and longevity, with signatures of adaptive evolution in multiple internal branches suggesting a step-wise evolution of some traits of interest. Most notably, we find multiple strands of evidence implicating selection events on genes affecting neural development and proliferation, in line with evidence of massive variation in the structure of Heliconiini brains (*10*, *11*, *29*). In conclusion, our work offers a comprehensive view to the evolutionary history of the tribe along with the architectures of genomic features that provide the first molecular explanation of physiological and behavioural innovations of a diverse group of butterflies.

Materials and Methods

DNA and RNA Extraction and Sequencing. Individuals of *Dryadula phaetusa*, *Dione juno*, *Agraulis vanilla vanillae*, were collected from partially inbred commercial stocks (Costa Rica Entomological Supplies, Alajuela, Costa Rica); individuals of *Agraulis vanilla incarnata* collected from Shady oak butterfly farm (Brooker, Florida, USA); while individuals of *Speyeria mormonia washingtonia* (Washington, USA), *Philaethria dido* (Gamboa, Panama), *Podothricha telesiphe* (Cocachimba, Peru), *H. aoede* (Tarapoto, Peru), *H. doris* (Gamboa, Panama), and *H. cydno*, were collected from the wild. Samples collected in Peru were obtained under permits 0289-2014-MINAGRI-DGFFS/DGEFFS, 020-014/GRSM/PEHCBM/DMA/ACR-CE, 040–2015/GRSM/PEHCBM/DMA/ACR-CE, granted to Dr Neil Rosser, and samples from Panama were collected under permits SEX/A-3-12, SE/A-7-13 and SE/AP-14-18. High-molecular-weight genomic DNA was extracted from pupae (commercial stock specimens) and adults (wild caught specimens), dissecting up to 100 mg of tissue, snap frozen in liquid nitrogen and homogenized in 9.2 ml buffer G2 (Qiagen Midi Prep Kit). The samples were then transferred to a 15 ml tube and processed with a Qiagen Midi Prep Kit (Qiagen, Valencia, CA) following the manufacturer's instructions. From the same stocks, RNA was extracted separately from six adult and early ommochrome stage pupae. Each tissue was frozen in liquid nitrogen and quickly homogenized in 500 µl Trizol, adding the remaining 500 µl Trizol at the end of the homogenization. Phase separation was performed by adding 200 µl of cold chloroform. The upper phase was then transferred to RNeasy Mini spin column and processed with a Qiagen RNeasy Mini Prep Kit (Qiagen, Valencia, CA), before DNAse purification using the Turbo DNA-free kit (Life Technologies, Carlsbad, CA) following the manufacturer's instructions. According to HWM genomic DNA quality, samples were sequenced with PacBio Circular Long Reads, PacBioHiFi, Oxford Nanopore Technology reads, or 10X Genomics Linked Reads, adding librarie s of Illumina DNAseq (HiSeq2500 150 x 2) data for PacBio Circular Long Reads libraries for error corrections. Short-read data for 39 other species where also downloaded from NCBI and used for *de novo* assembly and/or to curate already available assemblies (for further details see Supplementary Methods).

Long Read De Novo Genome Assembly. PacBio Hifi CCS reads were assembled using HIFIASM V0.12-r304 (*41*); while regular PacBio Circular Long Reads were corrected, trimmed and assembled using CANU V1.8 + 356 (*42*) as in Cicconardi et al. (2021). Resulted assemblies were subsequently corrected with their uncorrected raw PacBio long reads using PBMM2 V1.0.0 and ARROW V2.3.3. Further error correct was

performed with short Illumina reads using PILON v1.23 (*43*) with five iterations. Mis-assemblies corrected with POLAR STAR. Sequenced Illumina paired-end reads from 10X Genomics libraries were input to the SUPERNOVA v2.1.1 assembler (10x Genomics, San Francisco, CA, USA) (*44*) for *de novo* genome assembly, performing the optimisation of parameters. TIGMINT v1.1.2 (*45*) with default settings was then adopted to identify misassemblies. The final step of scaffolding was performed using ARCS v1.1.0 (*46*), a scaffolding procedure that utilizes the barcoding information contained in linked-reads to further organize assemblies. In all assemblies described thus far, haplocontigs were removed PURGE HAPLOTIGS v20191008 (*47*), and PacBio and Nanopore data (when available) were used to perform the first stage of scaffolding with LRSCAF v1.1.5 (*48*). If RNA-seq where available, P_RNA_SCAFFOLDER was also used to further scaffolding. Gap filling was performed with LR_GAPCLOSER v.1.1 (*49*). Before the chromosome-level scaffolding, we used synteny maps implemented with BLAST (*50*) and ALLMAPS (*51*) to identify duplicated regions at the end of scaffolds, manually curating the scaffolds to trim them away. Duplication level and completeness were checked with BUSCO (Benchmarking Universal Single-Copy Orthologs; v3.1.0, Insecta_odb9) (*52*) at each step of the assembling to keep track of losses and fragmentation of genomic regions.

Reference-based Genome Assembly. To assemble the genomes from the retrieved Illumina PE reads from NCBI (see Table S1) we implemented a reference-guided assembly approach, adapting and extending the protocol from Lischer and Shimizu (2017). The strategy involves first mapping reads against a reference genome of a related species (see TableS1 '*Ref Genome*' field) in order to reduce the complexity of the *de novo* assembly within continuous covered regions, then integrating reads with no similarity to the related genome in a further step. Modifications from the reference guided *de novo* pipeline (*53*) started in the 1st step, were paired-end illumina reads were mapped onto the reference genome with MINIMAP2 v2.17-r974-dirty (*54*), followed by PILON (*43*). This procedure the mapability of the target species to the reference by modifying the reference assembly at the nucleotide level. Paired-end reads were then mapped against the modified version of the reference assembly BOWTIE2 v2.2.1(*55*), and assigned into blocks as in the original pipeline. For each block reads were *de novo* assembled using SPADES v3.15 (*56*). Redundancy generated at this stage was removed as in the original pipeline. After the final step, short-reads were used to attempt scaffolding and gap closing with SOAPDENOVO2 VR240 (*57*). Leveraging the very small genetic differences between these species and their reference genomes, a final assembly scaffolding was performed with RAGOO (*58*), a homology-based scaffolding and misassembly correction pipeline. RAGOO (*58*) identifies structural variants and sequencing gaps, to accurately orders and orient *de novo* genome assemblies. ABYSS-SEALER v2.2.2 from the ABYSS package (*59*) was use as last step to attempt to close remaining gaps.

Curation of available Illumina Assemblies. Available Heliconiini assemblies from Edelman et al. (2019) were included in our dataset with a small, but effective curation. We checked for contaminants, as for the previous *de novo* and reference guided assemblies (see below), and haplocontig level (using BUSCO, see above. Raw Illumina reads were remapped onto their own assembly and PURGE HAPLOTIGS, with ad hoc -a parameter, was adopted to remove haplocontigs, followed by a scaffolding procedure with SOAPDENOVO2 (127mer). A synteny map was generated with ALLMAPS and, using their closest available reference assembly, a chromosomal scaffolding was generated. This procedure was adopted to maximise the contiguity. This represent a similar procedure recently adopted to scaffolded draft genomes (*61*). Finally, ABYSS-SEALER v2.2.2 from the ABYSS package (*59*) was used for gap filling (see above).

Bacterial Contamination & Assembly completeness assessment. After the genome assembly stage all datasets were analysed to remove contaminants. BLOBTOOLS v1.1.1 (*62*) was used to filter out any scaffold and contig assigned to fungal and bacterial contaminants. Furthermore, mitochondrial sequences were identified by blasting (BLASTN) contigs and scaffolds against the mitochondrial genome of available *Heliconius ssp*.. Finally a combination of BUSCO (*52*), with the Insecta set in ORTHODB v.9 [-m

genome], and EXONERATE v2.46.2 (*63*), was used to assess genome completeness and duplicated content.

Whole Genome Alignment & Genome evolution. BUSCO complete single-copy orthologous genes were used to generate a first draft of the phylogeny to guide the whole genome alignment. The nucleotide sequence of each locus aligned with MACSE v2.03 (*64*) and concatenated into a single alignment. A maximum-likelihood (ML) search was adopted to estimate the phylogenetic tree as implemented in FASTTREE v2.1.11 SSE3 (*65*). All 63 soft-masked genomes were then aligned with CACTUS v1.2.3 (*66*, *67*) with chromosome-level genomes as reference. Post-processing was performed by extracting information from the resulting hierarchical alignment (HAL). We then used HALSUMMARIZEMUTATIONS, from the CACTUS package, to summarize inferred mutations at each branch of the underlying Nymphalid phylogeny. We calculated rates for transposition (d_P), insertion (d_I), deletion (d_D), inversion (d_V), and duplication (d_U) events per million years (Ma) of evolution, based on the inferred divergence estimates from the phylogeny (see below). Ancestral state reconstruction of genome size was assessed using The maximum likelihood method implemented in the R package PHYTOOLS (*68*). Evolutionary conservation at individual alignment sites, PHYLOP scores (CONACC) were computed using a neutral model as implemented in HALPHYLOPTRAIN.PY script (CACTUS package). A non-overlapped sliding window of 10bp was adopted and data partitioned according to coding, intronic, 5'UTR and 3'UTR regions.

Gene Prediction and Transcriptome Annotation. The NCBI SRA archive was explored and the best SRA archives were downloaded, based on abundance and tissue (Table S2). These were used to re-annotate genes for their reference genomes (*e.g.*, *H. erato v. 1* and *H. melpomene v.2.5*). Short-reads were quality filtered and trimmed TRIMMOMATIC v0.39 (*69*), and predicted coding genes, *ab initio* and *de novo* approaches were implemented and combined in a pipeline with the aim of obtaining the best from each approach to overcome their own limitations. Reads from multiple tissues (when available) were pooled, mapped with STAR v2.7.10A (*70*) and used as training data for the BRAKER v2.1.5 pipeline (*71*), using AUGUSTUS v3.4.0 (*72*), along with the masked genomes generated with REPEATMASKER v4.1.1 (*73*), using the Lepidoptera database. To the prediction followed the UTR annotation step via GUSHR v1.1.0 (Gaius-Augustus/GUSHR, 2020). The *de novo* transcriptome assemblies were generated using TRINITY v2.10.0 (*74*, *75*) separately for each tissue. To generate the *ab initio* transcriptomes, tissue-specific reads were realigned to the genome using STARand assembled using both STRINGTIE v2.1.3B (*76*) and CUFFLINKS v2.2.1 (*77–79*). The BAM files were used as input for PORTCULLIS v1.1.2 (*80*) to validate splice-site DBs and together with the previously generate four types of annotations (prediction, *de novo*, two *ab initio*) were combined using MIKADO v2.3.3 (*81*). Finally, we used COMPARATIVE ANNOTATION TOOLKIT (CAT) (*82*), a comparative annotation pipeline that combines a variety of parameterizations of AUGUSTUS, including Comparative AUGUSTUS, with TRANSMAP projections, to annotate all the species through the whole-genome CACTUS alignments to produce an annotation set on every genome in alignment using *E. isabella* as a reference species.

Intron Evolution Analyses. Intronic regions were extracted from the longest transcript of each gene model using the annotations, as in Cicconardi et al 2021 (*16*). Sequences were scanned with REPEATMASKER. For each species introns were divided into short and long based on their median values. Their relative scaling coefficients and intercepts were subsequently analysed with SMATR (*83*). The intron turnover rate was subsequently estimated using MALIN (Mac OS X version) (*84*) to infer their conservation status in ancestral nodes, and the turnover rate (gain/loss) at each node with MALIN'S built-in model ML optimization procedure (1,000 bootstrap iterations).

Functional Annotation and Orthologous-Group Dynamic Evolution. Orthology inference was performed as implemented in BROCCOLI v1.1 (*85*) optimizing parameters (Romain Derelle, personal communication) for a more reliable list of single-copy orthologous groups (scOGs). BROCCOLI also returns a list of chimeric transcripts, which were manually curated in the *E. isabella* transcriptome, with a set

of custom script were implemented to automate the process in all the other taxa (BROCCOLICHIMERASPLITDATAGATHER.PY; available at https://francicco@bitbucket.org/ebablab/custum-scripts.git), before a second BROCCOLI iteration. From each OG a putative functional annotation was performed by identifying both the protein domain architecture using HMMER V3.3.2 (HMMSCAN) (*86*) with DAMA V2 (*87*). Annotation of GO terms was assigned with homology-based by searching against *Drosophila melanogaster* protein databases (FLYBASE.ORG), and with a predictive-based strategy with CATH assignments (*88–90*), scanning against the library of CATH functional family (FUNFAMS V4.2.0) HMMs (*90*). We modelled OG expansions and contractions as implemented in CAFE V5.0 using only genomes with complete BUSCO genes ≥ 90% (52/64 species).

Diversifying Positive Selection across Heliconiini. Evolutionary trajectories across Heliconiini were performed with a pipeline similar to that in Cicconardi *et al.*(*3*, *91*, *92*) computing ω (the ratio of nonsynonymous to synonymous substitution rates; d_N/d_S) on five branches of the Nymphalid phylogeny using codon-based alignments of groups of one-to-one orthologs (scOGs). Compared to previous pipelines, we introduced a pre-alignment filtering procedure as implemented in PREQUAL V1.02 (*93*), and a post-alignment filtering with HMMCLEANER (*94*). A ML gene tree was then generated as implemented in IQ-TREE2 V2.1.3 COVID-EDITION and the adaptive branch-site random effects likelihood (ABSREL) method (*95*, *96*) used, as implemented in the HYPHY batch language (*97*). Enrichment of GOTERMS was performed using a combination of two different approaches, the HYPERGTEST algorithm, implemented in the GOSTATS package (*98*) for R and GOATOOLS (*99*), and only considering significant terms in common between GOSTATS and GOATOOLS were considered (*P*-value < 0.05). The GENEMANIA prediction server (*100–102*) was used to predict functions of genes under selection (FDR cut-off of 0.05).

Phylogenetic Analysis & Divergence Estimates. Fully processed alignments of scOG were selected, concatenated and used to generate a maximum likelihood (ML) phylogenetic tree, as implemented in IQ-TREE2, partitioning the supermatrix for each locus and codon position, and with 5,000 ultrafast bootstrap replicates, resampling partitions, and then sites within resampled partitions (*103*, *104*). ILS was explored performing a coalescent summary method species tree, as implemented in ASTRAL-III v5.6.3 (*105*). To further explore phylogenetic support, the quartet sampling (QS) analysis was performed (*106*). The bayesian algorithm of MCMCTREE V4.8A (from PAML package) (*107*) with approximate likelihood computation was used to estimate divergence times for the whole dataset. Branch length were estimated by ML and then the gradient and Hessian matrix around these ML estimates were computed under MCMCTREE using the DNA supermatrix. From the TIMETREE database (*108*), four calibration points with uniform distributions were used (Table S3), consistent with previous phylogenetic studies of Heliconiini (*8*, *108*). For these priors a birth-death process with $\lambda=\mu=1$ and $\rho=0$, and a diffuse gamma-Dirichlet priors was given for the molecular rate ($\Gamma=2,20$) and a diffusion rate ($\sigma2=2,2$). Ten independent runs were executed, each with a burn-in of 2,500,000 generations. Convergence was checked using TRACER V1.7.1 (*109*).

Genome-wise Scan for Introgression. Patterns of introgression within Heliconiini were scanned using discordant-count test (DCT) and the branch-length test (BLT), which rely on the topologies of gene trees for triplets of species as implemented in Suvorov et al. (2021). These tests were applied on all triplets extracted from scOG gene trees within Heliconiini, and the resulting *P*-values were then corrected for multiple testing using the Benjamini-Hochberg procedure with a false discovery rate (FDR) cut-off of 0.05. DSUITE (*111*) was then used to plot the results in a heatmap plot (*110*).

Evolution of Gene Families. Fifty-seven gene families spanning receptors, enzymes, channels and transporters were selected to further explore the evolution of Heliconiini (Table S4). Amino acid sequences of the entire gene family (GF) were aligned using CLUSTALW v1.2.1 (*112*) and used to build a ML tree using FASTTREE V 2.1.11 SSE3 and used as input for MIPHY v1.1.2 (*113*), in order to automatically predict

members of orthologous groups for each GFs, leveraging a species tree. For each GF, OG copy number was processed with CAFE (see above) to further explore events of expansion and contraction. From each OG in-paralogs were removed (custom python script REMOVEINPARALOGFROMTREE.PY available at https://francicco@bitbucket.org/ebablab/custum-scripts.git). If the procedure generated a single-copy OG (nscOGs) it was analysed by contrasting evolutionary pressures between *Eueides* and *Heliconius* species. The signature of selection (aBSREL) and relaxation (RELAX (*114*)) were performed as implemented in HYPHY.

Cocoonases annotation & analysis. The protein sequences of *Heliconius* cocoonases were obtained from Smith et al. 2018 (*38*), while the sequence from *Bombyx mori* was downloaded from NCBI. Sequences were used as queries for map protein sequences onto all the 63 assemblies, using EXONERATE, and subsequently manually corrected. All nucleotide sequences were quality filtered using PREQUAL aligned using MACSE, to generated a single ML gene tree adopting IQ-TREE2. All sequences from each of the four clades were realigned separately and a number of tests implemented in HYPHY (overall ω; SLAC; ABSREL; RELAX). In particular, the sign of diversifying positive selection (ABSREL) was detected by scanning all internal branches of the whole cocoonases phylogeny, correcting for multiple tests using a final P-value threshold of 0.05. The structural analyses were conducted on the 181 full length sequences (~ 220 aa in length), removing peptide signal detected with SIGNALP v5.0b (*115*), and aligning sequences with the closest homolog protein for which a crystal structure is available (pdb: 4AG1), identified by HHPRED server (*116*). To predict 3D structures the ROSETTAFOLD v1.1.0 pipeline (*117*) was adopted. A graph theory based analysis was performed for each 3D model belonging to the final data set, as implemented in Ruiz-Serra et al. 2021 (*118*). The method adopts graph-based metrics in order to capture both the local features of the predicted distance maps (strength) as well as to characterize global patterns of the molecular interaction network. After performing a multiple alignment among all the sequences of the data set, we obtained an N x M output matrix, where N is the total number of the sequences and M the total number of all residue position, and used to performed a Principal Component Analysis (PCA) with the aim of projecting each M-dimensional vector (i.e., the set of strength values associated to each protein of the data set) into essential space (i.e., the PCA space). Consistency between the phylogenetic signal and the structural information of the four loci was evaluated by checking how groups are separated from each other in the PCA space, calculating four distributions of both the first and second component, and performing a Kolmogorov-Smirnov (K-S) test (*119*) as implemented in the R function KS.TEST. Key residues were identified selecting the residues that explained the most fraction of the first two PCA components.

REFERENCES AND NOTES

1. J. T. Stroud, J. B. Losos, Ecological Opportunity and Adaptive Radiation. *Annu. Rev. Ecol. Evol. Syst.* **47**, 507–532 (2016).
2. D. H. Erwin, A conceptual framework of evolutionary novelty and innovation. *Biol. Rev.* **96**, 1–15 (2021).
3. F. Cicconardi, P. Krapf, I. D'Annessa, A. Gamisch, H. C. Wagner, A. D. Nguyen, E. P. Economo, A. S. Mikheyev, B. Guénard, R. Grabherr, P. Andesner, A. Wolfgang, D. di Marino, F. M. Steiner, B. C. Schlick-Steiner, Genomic signature of shifts in selection in a sub-alpine ant and its physiological adaptations. *Mol. Biol. Evol.*, 1–17 (2020).
4. Y. Yuan, Y. Zhang, P. Zhang, C. Liu, J. Wang, H. Gao, A. Rus Hoelzel, I. Seim, M. Lv, M. Lin, L. Dong, H. Gao, Z. Yang, F. Caruso, W. Lin, R. R. Da Fonseca, D. Wang, X. Wang, M. H. Rasmussen, M. Liu, J. Zheng, L. Zhao, P. F. Campos, H. Kang, M. Iversen, Y. Song, X. Guo, J. Guo, Y. Qin, S. Pan, Q. Xu, L. Meng, A. Yunga, S. Liu, S. Ming-Yuen Lee, X. Liu, X. Xu, H. Yang, G. Fan, K. Wang, S. Li, Comparative genomics provides insights into the aquatic adaptations of mammals. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–9 (2021).
5. K. M. Kozak, M. Joron, W. O. McMillan, C. D. Jiggins, Rampant Genome-Wide Admixture across the Heliconius Radiation. *Genome Biol. Evol.* **13**, 1–17 (2021).
6. S. H. Martin, J. W. Davey, C. Salazar, C. D. Jiggins, Recombination rate variation shapes barriers to introgression across butterfly genomes.

PLoS Biol. (2019), doi:10.1371/journal.pbio.2006288.

7. N. B. Edelman, P. B. Frandsen, M. Miyagi, B. Clavijo, J. Davey, R. B. Dikow, G. García-accinelli, S. M. Van Belleghem, N. Patterson, Genomic architecture and introgression shape a butterfly radiation. **599**, 594–599 (2019).
8. K. M. Kozak, N. Wahlberg, A. F. E. Neild, K. K. Dasmahapatra, J. Mallet, C. D. Jiggins, Multilocus species trees show the recent adaptive radiation of the mimetic heliconius butterflies. *Syst. Biol.* **64**, 505–524 (2015).
9. É. C. P. de Castro, J. P. Zurano, É. C. Pinheiro de Castro, M. Zagrobelny, J. P. Zurano, M. Zikan Cardoso, R. Feyereisen, S. Bak, Sequestration and biosynthesis of cyanogenic glucosides in passion vine butterflies and consequences for the diversification of their host plants. *Ecol. Evol.* **9**, 5079–5093 (2019).
10. F. J. Young, S. H. Montgomery, Pollen feeding in Heliconius butterflies : the singular evolution of an adaptive suite. *Proc. R. Soc. B Biol. Sci.* **287** (2020), doi:10.1098/rspb.2020.1304.
11. S. H. Montgomery, R. M. Merrill, S. R. Ott, Brain composition in Heliconius butterflies, posteclosion growth and experience-dependent neuropil plasticity. *J. Comp. Neurol.* **524**, 1747–1769 (2016).
12. J. R. Walters, C. Corbins, T. J. Hardcastle, C. D. Jiggins, Evaluating female remating rates in light of spermatophore degradation in Heliconius butterflies: Pupal-mating monandry versus adult-mating polyandry. *Ecol. Entomol.* **37**, 257–268 (2012).
13. T. J. Thurman, E. Brodie, E. Evans, W. O. McMillan, Facultative pupal mating in Heliconius erato: Implications for mate choice, female preference, and speciation. *Ecol. Evol.* **8**, 1882–1889 (2018).
14. A. Kapusta, A. Suh, C. Feschotte, Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U. S. A.* (2017), doi:10.1073/pnas.1616702114.
15. A. A. Ruggieri, L. Livraghi, J. J. Lewis, E. Evans, F. Cicconardi, L. Hebberecht, S. H. Montgomery, A. Ghezzi, J. A. Rodriguez-Martinez, C. D. Jiggins, W. O. McMillan, B. A. Counterman, R. Papa, S. M. Van Belleghem, *bioRxiv*, in press (available at https://www.biorxiv.org/content/10.1101/2022.04.14.488334v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.04.14.488334v1.abstract).
16. F. Cicconardi, J. J. Lewis, S. H. Martin, R. D. Reed, C. G. Danko, S. H. Montgomery, Chromosome Fusion Affects Genetic Diversity and Evolutionary Turnover of Functional Loci but Consistently Depends on Chromosome Size. *Mol. Biol. Evol.* **38**, 4449–4462 (2021).
17. C. Sun, J. Huang, Y. Wang, X. Zhao, L. Su, G. W. C. Thomas, M. Zhao, X. Zhang, I. Jungreis, M. Kellis, S. Vicario, I. V. Sharakhov, S. M. Bondarenko, M. Hasselmann, C. N. Kim, B. Paten, L. Penso-Dolfin, L. Wang, Y. Chang, Q. Gao, L. Ma, L. Ma, Z. Zhang, H. Zhang, H. Zhang, L. Ruzzante, H. M. Robertson, Y. Zhu, Y. Liu, H. Yang, L. Ding, Q. Wang, D. Ma, W. Xu, C. Liang, M. W. Itgen, L. Mee, G. Cao, Z. Zhang, B. M. Sadd, M. W. Hahn, S. Schaack, S. M. Barribeau, P. H. Williams, R. M. Waterhouse, R. L. Mueller, Genus-Wide Characterization of Bumblebee Genomes Provides Insights into Their Evolution and Variation in Ecological and Behavioral Traits. *Mol. Biol. Evol.* **38**, 486–501 (2021).
18. D. E. Neafsey, R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, J. E. Allen, J. Amon, B. Arcà, P. Arensburger, G. Artemov, L. A. Assour, H. Basseri, A. Berlin, B. W. Birren, S. A. Blandin, A. I. Brockman, T. R. Burkot, A. Burt, C. S. Chan, C. Chauve, J. C. Chiu, M. Christensen, C. Costantini, V. L. M. Davidson, E. Deligianni, T. Dottorini, V. Dritsou, S. B. Gabriel, W. M. Guelbeogo, A. B. Hall, M. V. Han, T. Hlaing, D. S. T. Hughes, A. M. Jenkins, X. Jiang, I. Jungreis, E. G. Kakani, M. Kamali, P. Kemppainen, R. C. Kennedy, I. K. Kirmitzoglou, L. L. Koekemoer, N. Laban, N. Langridge, M. K. N. Lawniczak, M. Lirakis, N. F. Lobo, E. Lowy, R. M. MacCallum, C. Mao, G. Maslen, C. Mbogo, J. McCarthy, K. Michel, S. N. Mitchell, W. Moore, K. A. Murphy, A. N. Naumenko, T. Nolan, E. M. Novoa, S. O'Loughlin, C. Oringanje, M. A. Oshaghi, N. Pakpour, P. A. Papathanos, A. N. Peery, M. Povelones, A. Prakash, D. P. Price, A. Rajaraman, L. J. Reimer, D. C. Rinker, A. Rokas, T. L. Russell, N. Sagnon, M. V. Sharakhova, T. Shea, F. A. Simão, F. Simard, M. A. Slotman, P. Somboon, V. Stegniy, C. J. Struchiner, G. W. C. Thomas, M. Tojo, P. Topalis, J. M. C. Tubio, M. F. Unger, J. Vontas, C. Walton, C. S. Wilding, J. H. Willis, Y. C. Wu, G. Yan, E. M. Zdobnov, X. Zhou, F. Catteruccia, G. K. Christophides, F. H. Collins, R. S. Cornman, A. Crisanti, M. J. Donnelly, S. J. Emrich, M. C. Fontaine, W. Gelbart, M. W. Hahn, I. A. Hansen, P. I. Howell, F. C. Kafatos, M. Kellis, D. Lawson, C. Louis, S. Luckhart, M. A. T. Muskavitch, J. M. Ribeiro, M. A. Riehle, I. V. Sharakhov, Z. Tu, L. J. Zwiebel, N. J. Besansky, Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science (80-.).* **347** (2015), doi:10.1126/science.1258522.
19. J. L. Da Lage, G. W. C. Thomas, M. Bonneau, V. Courtier-Orgogozo, Evolution of salivary glue genes in Drosophila species. *BMC Evol. Biol.* **9** (2018), doi:10.1101/359190.
20. É. C. Pinheiro de Castro, R. Demirtas, A. Orteu, C. E. Olsen, M. S. Motawie, M. Zikan Cardoso, M. Zagrobelny, S. Bak, The dynamics of cyanide defences in the life cycle of an aposematic butterfly: Biosynthesis versus sequestration. *Insect Biochem. Mol. Biol.* **116**, 103259 (2020).
21. S. E. W. Opitz, C. Müller, Plant chemistry and insect sequestration. *Chemoecology*. **19**, 117–154 (2009).
22. M. Du, X. Yin, S. Zhang, B. Zhu, Q. Song, S. An, Identification of lipases involved in PBAN stimulated Pheromone production in Bombyx mori using the DGE and RNAi approaches. *PLoS One*. **7** (2012), doi:10.1371/journal.pone.0031045.
23. E. J. Sung, M. Ryuda, H. Matsumoto, O. Uryu, M. Ochiai, M. E. Cook, N. Y. Yi, H. Wang, J. W. Putney, G. S. Bird, S. B. Shears, Y. Hayakawa, Cytokine signaling through Drosophila Mthl10 ties lifespan to environmental stress. *Proc. Natl. Acad. Sci. U. S. A.* (2017), doi:10.1073/pnas.1712453115.
24. L. Wu, Q. Jia, X. X. Zhang, X. X. Zhang, S. Liu, Y. Park, R. Feyereisen, K. Y. Zhu, E. Ma, J. Zhang, S. Li, CYP303A1 has a conserved function in adult eclosion in Locusta migratoria and Drosophila melanogaster. *Insect Biochem. Mol. Biol.* **113**, 103210 (2019).
25. B. Tang, S. Wang, F. Zhang, Two storage hexamerins from the beet armyworm Spodoptera exigua: Cloning, characterization and the effect of gene silencing on survival. *BMC Mol. Biol.* **11** (2010), doi:10.1186/1471-2199-11-65.
26. P. Portin, P. Portin, General outlines of the molecular genetics of the

Notch signalling pathway in Drosophila melanogaster: A review. *Hereditas*. **136**, 89–96 (2002).
27. X. Li, Y. Xie, S. Zhu, Notch maintains Drosophila type II neuroblasts by suppressing expression of the fez transcription factor earmuff. *Dev.* **143**, 2511–2521 (2016).
28. S. M. Farris, Evolution of complex higher brain centers and behaviors: Behavioral correlates of mushroom body elaboration in insects. *Brain. Behav. Evol.* **82**, 9–18 (2013).
29. A. Couto, F. Young, M. Stephen, Muschroom body expantion in Heliconiini. *prep* (2023).
30. M. R. Sahu, A. C. Mondal, Neuronal Hippo signaling: From development to diseases. *Dev. Neurobiol.* **81**, 92–109 (2021).
31. A. Kaya-çopur, F. Marchiano, M. Y. Hein, D. Alpern, J. Russeil, N. M. Luis, M. Mann, B. Deplancke, B. H. Habermann, F. Schnorrer, The hippo pathway controls myofibril assembly and muscle fiber growth by regulating sarcomeric gene expression. *Elife*. **10**, 1–34 (2021).
32. N. Abeysundara, A. J. Simmonds, S. C. Hughes, Moesin is involved in polarity maintenance and cortical remodeling during asymmetric cell division. *Mol. Biol. Cell*. **29**, 419–434 (2018).
33. X. Wang, Y. Zhang, S. S. Blair, Fat-regulated adaptor protein Dlish binds the growth suppressor Expanded and controls its stability and ubiquitination. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1319–1324 (2019).
34. S. Bahrampour, S. Thor, Ctr9, a key component of the paf1 complex, affects proliferation and terminal differentiation in the developing drosophila nervous system. *G3 Genes, Genomes, Genet.* **6**, 3229–3239 (2016).
35. N. Loyer, J. Januschke, Where does asymmetry come from? Illustrating principles of polarity and asymmetry establishment in Drosophila neuroblasts. *Curr. Opin. Cell Biol.* **62**, 70–77 (2020).
36. S. Blair, H. McNeill, Big roles for Fat cadherins. *Curr. Opin. Cell Biol.* **51**, 73–80 (2018).
37. K. Yildirim, J. Petri, R. Kottmeier, C. Klämbt, Drosophila glia: Few cell types and many conserved functions. *Glia*. **67**, 5–26 (2019).
38. G. Smith, J. E. Kelly, A. Macias-Muñoz, C. T. Butts, R. W. Martin, A. D. Briscoe, Evolutionary and structural analyses uncover a role for solvent interactions in the diversification of cocoonases in butterflies. *Proc. R. Soc. B Biol. Sci.* **285** (2018), doi:10.1098/rspb.2017.2037.
39. T. Gai, X. Tong, M. Han, C. Li, C. Fang, Y. Zou, H. Hu, H. Xiang, Z. Xiang, C. Lu, F. Dai, Cocoonase is indispensable for Lepidoptera insects breaking the sealed cocoon. *PLoS Genet.* **16**, 1–16 (2020).
40. L. Gerasimavicius, B. J. Livesey, J. A. Marsh, Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure, 1–15 (2022).

ACKNOWLEDGMENTS

This article would not possible without the great support of the great *Heliconius* community. We are also grateful to the environmental ministries in Peru and Panama for permission to collect and export samples and the STRI community for assistance in the field. F.C. would like to thank Ronald Mori Pezo for his great help in collecting *H. aoede* and *Podotricha telesiphe*; Angel Corpuz for various informatics support including a great patience; Ian Fiddes, Mark Diekhans, Glenn Hichey and Marina Haukness for their great support for Cactus and CAT; Federica Cattonaro, Davide Scaglione and Simone Scalabrin from IGA (Udine, Italy) for their fruitful discussion on the best sequencing strategy to perfom. F.C. and S.H.M. are grateful to the High-Performance Computing team at the Advanced Computing Research Centre, University of Bristol for support. E.P.C. was supported by the H2020 Marie Skłodowska-Curie Actions (Cyanide Evolution/841230).

Funding: This article was supported by NERC IRF (NE/N014936/1) and ERC Starter Grant (758508) to S.H.M. S.M. was supported by the Royal Society URF\R1\180682. F.C. was supported as a postdoctoral researcher ERC. **Author contributions:** Conceptualization: S.H.M and F.C. Data collection: All authors. Genomic analysis: F.C. Structural analyses: F.C., E.M. and D.d.M. Visualization: F.C. Funding acquisition: S.H.M. Writing – original draft: F.C. and S.H.M. Writing – review & editing: All authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data and code used for these analyses are available on NCBI and GitHub. Individual mark-recapture datasets can be obtained by contacting specific dataset owners.
License information: Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/ about/science-licenses-journal-article-reuse